Progression of Live Operation Monitoring

David Huff (New Mexico Institute of Mining and Technology)

Mentor: Daniel Illescas(HPC-ENV), Mike Mason(HPC-ENV)

Background

The Operations(Ops) team uses a Splunk Application to augment their ability to react and fix the multiple LANL clusters. Splunk is a messages analytic web application that can take in computer logs and search them for useful info. In this ever evolving field, it is important for this app to keep up with the new systems, architecture, and monitoring tools being deployed within HPC.

/ebsites								
dar On-call Tech Ops Wiki			Rec	i=Page on-call Yellow=e	mail Blue=DST			
Services		low		Campaign			Logs monitoring Su9	
		Fenrir monitoring 7x24						
low	low	low	low	low	low	low	low	
ba 95.91% monitoring 5x9	CP 0.57%	fg 6.06%	GI 98.32% monitoring 5x9	KO 72.18%	Sn 96.2% monitoring 5x9	21.50% monitoring 5x9	WC 92.61% monitoring 5x9	
ork Status		hw_logger (Last Week)						
\$	Status ÷	cluster time time	message ≑					
r	severe	grizzly 2019-07-15 09:12:57		ESEAT: The network card has been reseated do t	to no link. Node is still not linking up. Conti	nuing to troubleshoot:robbieg		
Ly	low low severe	cyclone 2019-07-15 08:26:21						
· · ·	low	grizzly 2019-07-12 15:26:35	CABLE:gr0114-e4p36::RESEAT:Cable was throwing local 1			.:dustin_newt,JJ		
	low	grizzly 2019-07-12 15:25:38	CABLE:gr0331-e10p27::REPLACE:replaced_part_number_and outSN='WY6045F' part_from='Grizzly Spare Parts':dus		rt_number='' problem_description='Cable has b	een consistently throwing errors and has	s been replaced.' inSN='S5185A3002W'	
		grizzly 2019-07-11 09:49:38	NODE:gr0692:bash: dmidecode: command not found:CADDY_	SLAM:Node is not responding, caddy slammed to	o see if issue has been resolved. Further troub	leshooting will take place.:dustin_newt		
		kodiak 2019-07-11 09:47:31	NODE:ko029::REPLACE:replaced_part_number_and_descript delivery':robbieg	ion='New node' replacement_part_number=''	problem_description='Node was not able to see	a bulk of its hardware' inSN='p100208	8333' outSN='p100165627' part_from='	
		badger 2019-07-11 09:45:20	NODE:ba607::REPLACE:replaced_part_number_and_descript outSN='p100189034' part_from='Penguin Delivery':rob		problem_description='Node was unresposive aft	er multiple HW reseats as well as new ca	able replacments.' inSN='p100269397'	
		kodiak 2019-07-11 08:49:44	NODE:ko128:P100196150:REPLACE:replaced_part_number_an part_from='Kodiak parts cabinet':robbieg	d_description='32g DIMM' replacement_part_r	number='' problem_description='Node was throw	ing uncorrectable memory error of boot u	up.' inSN='2g3b1ll' outSN='2g3b1rl'	
		badger 2019-07-10 14:55:15	NODE:ba212:P100167838:OTHER:LinkDowned. Replaced cabl	e.:jurneyjesse,DN,BC			« prev 1 2 3 4	

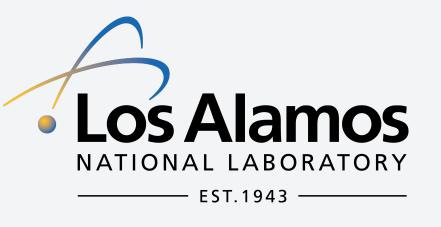
Fig. (1): This is the home dashboard the Ops team use on a daily basis

Acknowledge Script

Currently there are two Splunk search heads in production. One is for the general use of the entire HPC staff, while the second is set aside specifically for the Ops team. Although the other teams can see what errors occurred, due to the separation, they do not know what action the Ops team have taken to correct the errors. To fix this a script was implemented to sync the two databases together. The script takes the acknowledged events from a local Splunk KV Store, stored in a MongoDB, and transfers it over to an identical database on the other search head. This was done in the hopes that it would increase transiency and communication for the Ops team and cluster admins. A new Splunk cluster is also being created to improve scalability and to remove the need to sync databases.

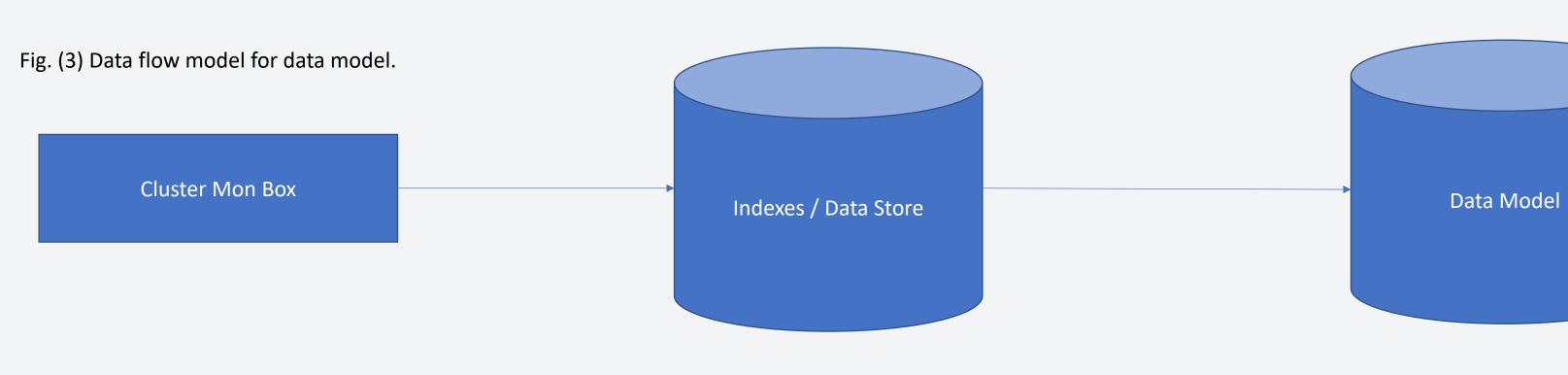
Acknowledged Events												
Original Event Time \$	Acknowledged At \$	Acknowledged Until \$	device \$	type ≑	message 🕏	raw ≑	notes	username				
2019-07- 11 08:29:58	2019-07-11 08:38:32	2019-07-11 13:38:32	core3 Port 26	LocalLinkIntegrityErrors	<pre>[LocalLinkIntegrityErrors == 4194397] - (edge26 Port 3)</pre>	<14>Jul 11 08:29:58 gr-master hsnmon[126901]: core3 Port 26: [LocalLinkIntegrityErrors == 4194397] - (edge26 Port 3)						
2019-07- 11 08:35:22	2019-07-11 08:42:45	2019-07-12 08:42:45	gr1392	pingStatus	compute node gr1392 is down	<13>Jul 11 08:35:22 gr-mon2.lanl.gov pingStatus[111437]: compute node gr1392 is down						
2019-07- 11 08:35:22	2019-07-11 08:42:45	2019-07-12 08:42:45	gr0843	pingStatus	compute node gr0843 is down	<13>Jul 11 08:35:22 gr-mon2.lanl.gov pingStatus[111436]: compute node gr0843 is down						
2019-07- 11 08:29:58	2019-07-11 08:37:41	2019-07-12 08:37:41	core3 Port 26	LocalLinkIntegrityErrors	<pre>[LocalLinkIntegrityErrors == 4194397] - (edge26 Port 3)</pre>	<14>Jul 11 08:29:58 gr-master hsnmon[126901]: core3 Port 26: [LocalLinkIntegrityErrors == 4194397] - (edge26 Port 3)						
2019-07- 11 09:02:24	2019-07-11 09:21:01	2019-07-12 09:21:01	gr0646 edge21-18 1M CU Missing Link	Check Cable	Check Cable: gr0646 edge21-18 1M CU Missing Link	<14>Jul 11 09:02:24 gr-master hsnmon[133977]: Check Cable: gr0646 edge21-18 1M CU Missing Link						

Fig. (2): example of what the Ops team deal with on a daily basis



Data Model Change

One of the benefits of Splunk is the ability to create schemas on the fly. This is generally beneficial when exploring your data but we found doing this repeatedly over large data sets caused performance issues. To prevent an overload, a data model was implemented to store and structure error messages explained in figure 3. The new data model is modular, which allows for easy changes to the model when new systems are added or removed. The model also speeds up the search process allowing the Ops team better utilize their time.



NVIDIA GPU Monitoring

With new technology comes new problems. Kodiak is a GPU cluster, however, it was found that some of the GPUs on the system were staying in a constant throttled state causing jobs to run slow. By running nvidia-smi to sample stats on Kodiak's compute nodes, we can log this data into the monitoring infrastructure. Allowing us to alert Ops when this occurs so they can take appropriate actions. The data also gives another dimension to study for future analysis.

Future Work: Statistical Analysis

Syslog

Currently we are testing out new Splunk libraries. Although, it is currently unclear if the new functionality will be implemented into a dashboard. We are using basic statistical analysis to make simple job and system profiles. The images below show both types of profiles. Figure 4 tabulates the average number of syslog messages the cluster produces every hour and the second standard deviation. This information can be used to perform regression analysis with a confidence interval of 95%. The next table compares a single job on one node against the cluster statistics to find outliers. The time charts in figure 5 are Splunk's equivalent to the regression analysis. The top time chart is the number of syslog messages produced every hour and the blue shaded area is the confidence interval.

Cooling Distribution Unit(CDU)

The bottom time chart in figure 5 is the same process but uses data collected from the CDU. With new Splunk Analytic tools we hope to improve the current CDU alerts. The current alerts often report false alarms caused by intermittent outliers from the sensors, and have difficulties detecting sensor problems vs actual CDU problems. With smarter CDU monitoring it will be easier for the Ops team to identify problems before any damage happens on the system.



Fig. (4): This image depicts the the confidence interval and how we are using it to to compare syslog messages.

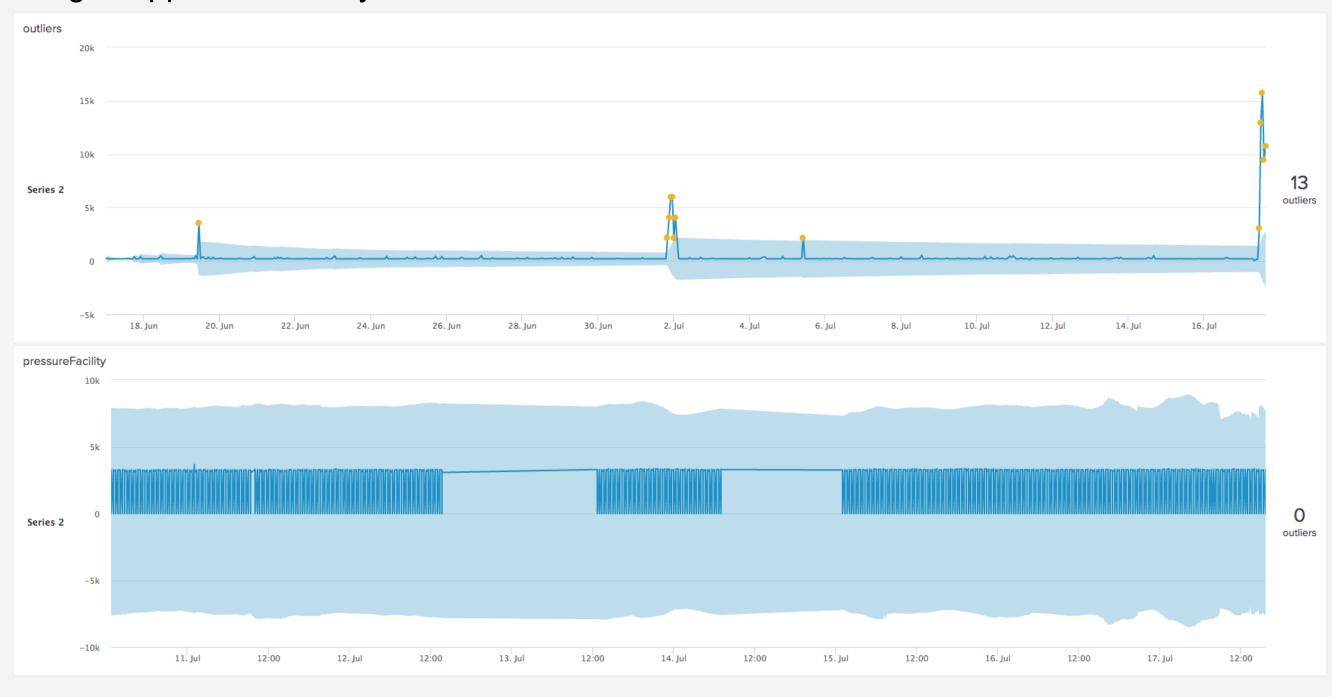


Fig. (5): The time charts are Splunk's version of regression analysis applied to syslog messages(top) and Cooling Distribution Unit(bottom)

